Animal (2016), **10:6**, pp 1067–1075
doi:10.1017/S1751731115001792

animal

Sharing reference data and including cows in the reference population improve genomic predictions in Danish Jersey

G. Su[1†], P. Ma[1], U. S. Nielsen[2], G. P. Aamand[3], G. Wiggans[4], B. Guldbrandtsen[1] and M. S. Lund[1]

[1]*Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, DK-8830 Tjele, Denmark;* [2]*Seges, DK-8200 Aarhus, Denmark;* [3]*Nordic Cattle Genetic Evaluation, DK-8200 Aarhus, Denmark;* [4]*Agricultural Research Service, USDA, Beltsville, MD 20705-2350, USA*

(Received 5 April 2015; Accepted 4 July 2015; First published online 2 September 2015)

Small reference populations limit the accuracy of genomic prediction in numerically small breeds, such like Danish Jersey. The objective of this study was to investigate two approaches to improve genomic prediction by increasing size of reference population in Danish Jersey. The first approach was to include North American Jersey bulls in Danish Jersey reference population. The second was to genotype cows and use them as reference animals. The validation of genomic prediction was carried out on bulls and cows, respectively. In validation on bulls, about 300 Danish bulls (depending on traits) born in 2005 and later were used as validation data, and the reference populations were: (1) about 1050 Danish bulls, (2) about 1050 Danish bulls and about 1150 US bulls. In validation on cows, about 3000 Danish cows from 87 young half-sib families were used as validation data, and the reference populations were: (1) about 1250 Danish bulls, (2) about 1250 Danish bulls and about 1150 US bulls, (3) about 1250 Danish bulls and about 4800 cows, (4) about 1250 Danish bulls, 1150 US bulls and 4800 Danish cows. Genomic best linear unbiased prediction model was used to predict breeding values. De-regressed proofs were used as response variables. In the validation on bulls for eight traits, the joint DK-US bull reference population led to higher reliability of genomic prediction than the DK bull reference population for six traits, but not for fertility and longevity. Averaged over the eight traits, the gain was 3 percentage points. In the validation on cows for six traits (fertility and longevity were not available), the gain from inclusion of US bull in reference population was 6.6 percentage points in average over the six traits, and the gain from inclusion of cows was 8.2 percentage points. However, the gains from cows and US bulls were not accumulative. The total gain of including both US bulls and Danish cows was 10.5 percentage points. The results indicate that sharing reference data and including cows in reference population are efficient approaches to increase reliability of genomic prediction. Therefore, genomic selection is promising for numerically small population.

Keywords: genomic selection, reliability, Jersey cattle, reference population

Implications

Reference population size is a key factor affecting accuracy of genomic prediction. In dairy cattle, reference population usually consisted of progeny-tested bulls. Limited number of progeny-tested bulls is a limitation to the accuracy of genomic prediction in numerically small breeds. The results from this study indicate that sharing reference data and including cows in reference population can greatly increase reliability of genomic prediction for the populations where the size of domestic bull reference population is small.

† E-mail: guosheng.su@mbg.au.dk

Introduction

Genomic selection has been widely implemented in dairy cattle breeding. Its success depends on accurate genomic predictions. A key factor affecting accuracy of genomic prediction is the amount of information from reference population (Daetwyler *et al.*, 2008; Goddard, 2009; Goddard and Hayes, 2009). In dairy cattle, reference populations are usually composed of progeny-tested bulls, since they have reliable phenotypic information from a large group of daughters. However, the number of progeny-tested bulls is limited for numerically small dairy cattle populations, such as Danish Jersey.

Currently, there are about 60 000 cows in the Danish Jersey population. Until now, only about 1200 to 1400

Danish progeny-tested bulls (depending on trait) are available to be used as reference bulls. Due to the small reference population, accuracy of genomic prediction in the Danish Jersey is much lower than in the Danish Holstein and Red Cattle populations (Su *et al*., 2011 and 2012c; Gao *et al*., 2012; Thomasen *et al*., 2012). Therefore, it is important to find efficient approaches to improve accuracy of genomic prediction in this population.

Several approaches have been proposed to improve accuracy of genomic prediction for small dairy cattle populations (Lund *et al*., 2014). An efficient approach is to use a joint reference population that combines the reference data from different populations. A large benefit from this approach has been reported in genomic prediction for North American Holstein populations (Schenkel *et al*., 2009; Muir *et al*., 2010), European Holstein populations (Lund *et al*., 2011), Chinese Holstein population (Zhou *et al*., 2013) and Brown Swiss populations (VanRaden *et al*., 2012). However, since accuracy of genomic prediction depends on the relationship between candidates and reference animals (Lund *et al*., 2009; Habier *et al*., 2010; Clark *et al*., 2012; Pszczola *et al*., 2012), it requires that the reference populations are close enough to link with the target populations. Another approach is to genotype cows and include them in the reference population. Cow reference populations have been used to predict genomic breeding values in populations where only few progeny-tested bulls have reliable phenotypic information (Ding *et al*., 2013; Li *et al*., 2014). A more common approach is to include genotyped cows in a bull reference population. Increasing the accuracy of genomic prediction by adding cows to a progeny-tested bull reference population has been reported in previous studies (Wiggans *et al*., 2010; Calus *et al*., 2013; Cooper *et al*., 2015). Although phenotypic information is much less accurate for cows than progeny-tested bulls, the increased information may be considerable because a large number of cows are available to be used as reference animals. According to previous studies (Daetwyler *et al*., 2008; Goddard and Hayes, 2009), the gain from additional information depends on the size of original reference population.

Both approaches can be implemented to improve accuracy of genomic prediction for Danish Jersey. On one hand, US Jersey has been contributed to the Danish Jersey population for a long time, especially during the period from 1985 to 1995. It is expected that a joint reference population combining Danish and US Jersey bulls would increase accuracy of genomic prediction considerably for both Danish and US Jersey populations. Therefore, in 2013, marker data for Danish and US Jersey bulls were exchanged to create a joint reference population for genomic prediction of Jersey cattle. On another hand, adding cows to the reference population of Danish Jersey may increase accuracy of genomic prediction considerably, because the current progeny-tested bull reference population is small. Taking this into consideration, a number of females have been genotyped since 2013 with the purpose of increasing the size of the reference population.

The objective of this study was to investigate the improvement of genomic predictions in numerically small breeds by sharing reference data and including cows in reference population. Thus, this study accessed the reliability and unbiasedness of genomic breeding values predicted using a Danish Jersey bull reference population, a joint Danish–US Jersey bull reference population, a reference population consisting of Danish Jersey bulls and cows, and a reference population consisting of Danish Jersey bulls and cows as well as US Jersey bulls. The validation was carried out on Danish Jersey bulls and cows, respectively.

Material and methods

Data

The data in the analysis included 1369 Danish Jersey bulls, 1160 US Jersey bulls and 9419 Danish Jersey cows; 98.4% of Danish bulls were born from 1988 to 2010, 99.6% of US bulls from 2000 to 2009, and 95.4% of Danish cows from 2010 to 2013. The cows were from herds with good data registration. Danish Jersey bulls were genotyped with the Illumina Bovine SNP50 chip (54 609 single nucleotide polymorphism (SNP, Illumina, Inc.)). US Jersey bulls were genotyped either with the standard Illumina Bovine SNP50 chip or with the GeneSeek Genomic Profiler chip HD (near 78 000 SNP, GeneSeek, Neogen Corporation). Danish Jersey cows were genotyped either with the standard BovineLD BeadChip (6909 SNP, Illumina, Inc.) or with a customized Illumina BovineLD which included the SNP in the standard BovineLD BeadChip and near 5000 user-selected SNP, and a few cows were genotyped with Illumina Bovine SNP50 chip. The marker data of different chips were imputed to Bovine SNP50 chip using FIMPUTE (Sargolzaei *et al*., 2014). The markers which are not in the Bovine SNP50 chip were excluded. After removing markers with allele frequency <1%, 39 937 autosomal markers were used to predict genomic breeding values.

De-regressed proofs (DRP) derived from the published estimated breeding values (EBV) of the Interbull 2014-12 evaluation and Nordic 2015-02 evaluation were used as response variables in the analysis. Two set of DRP were obtained. One (DRP_B) was derived from a de-regression procedure in which EBV of genotyped Danish bulls and US bulls were used. The other (DRP_{BC}) was derived based on EBV of all genotyped animals including cows. The traits with reliable DRP for both Danish and US Bulls were milk, fat, protein, body conformation, fertility, longevity, mastitis and udder conformation. The traits with DRP available for Danish bulls, US Bulls and genotyped Danish cows were milk, fat, protein, body conformation, mastitis and udder conformation. Thus, eight traits were analyzed when using DRP_B, and six traits when using DRP_{BC}.

Validation of genomic predictions

Genomic estimated breeding values (GEBV) using different data sets were validated on bulls and cows, respectively. The bull validation set comprised Danish Jersey bulls born in the years from 2005 onwards, which accounted for about 25% of

https://doi.org/10.1017/S1751731115001792 Published online by Cambridge University Press

the Danish Jersey bulls. This validation set was suitable for comparing GEBV from Danish bull reference and the joint Danish and US reference populations. However, it is not appropriate to use this validation set to assess genomic predictions based on the reference population including cows, since most cows in the reference population were born during 2010-12 and were the sibs or daughters of the bulls in the bull validation set. Therefore, a cow validation set was created in the following way: (1) genotyped cows born in the period from 1 July 2012 onwards were extracted; (2) these cows and their paternal female half-sibs born after 2007 were defined as the cow validation set but excluding the half-sib families with size >500 (these families were kept in the reference data in order to avoid a large reduction of reference population size). This resulted in about 3000 validation cows from 87 paternal half-sib families. When using the cow validation set, validation cows' maternal female and male half-sibs born after 2007 were excluded from reference population. In addition, the progenies of these animals (validation cows and the sibs) were also removed from the reference population. The validation and reference populations defined this way were in order to reduce the relationship between validation and reference animals and to achieve consistency with a real life selection scenario as much as possible.

For the bull validation set, the reference populations were: (1) Danish bull reference population, and (2) the Joint Danish and US bull reference population. For the cow validation set, four reference populations were used for genomic prediction: (1) Danish bulls, (2) Danish bulls and US bulls, (3) Danish bulls and cows and (4) Danish bulls, US bulls and Danish cows. The validation scenarios, the number of animals in the reference data and the validation data are shown in Table 1. Genomic predictions using different reference data sets were evaluated by comparing GEBV and DRP for animals in the validation data. Reliability of GEBV was measured as the squared correlation between GEBV and DRP divided by the reliability of DRP (Su *et al*., 2012b). Unbiasedness of genomic prediction was assessed by regression of DRP on GEBV (Su *et al*., 2012a).

Prediction model

Breeding values were predicted using a genomic best linear unbiased model (GBLUP), based on different reference populations. The GBLUP model is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}_g\mathbf{g} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of DRP, μ the overall mean, $\mathbf{1}$ a vector of ones, $\mathbf{g}$ the vector of additive genomic effects, $\mathbf{Z}_g$ the incidence matrix linking $\mathbf{g}$ to $\mathbf{y}$ and $\mathbf{e}$ the vector of residuals.

It is assumed that $\mathbf{g} \sim N(\mathbf{0}, \mathbf{G_A}\sigma^2{}_g)$, and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{D}\sigma^2{}_e)$, where $\mathbf{G_A}$ is a genomic relationship matrix combining marker-based relationship matrix and pedigree-based relationship matrix, $\sigma^2{}_g$ is the additive genetic variance, $\mathbf{D}$ is a diagonal matrix and $\sigma^2{}_e$ is the residual variance. Matrix $\mathbf{D}$ has diagonal elements $d_{ii} = (1 - r^2{}_{\mathrm{DRP}})/r^2{}_{\mathrm{DRP}}$ to account for heterogeneous residual variances due to different reliabilities of DRP ($r^2{}_{\mathrm{DRP}}$).

The matrix $\mathbf{G_A}$ was constructed in the following steps. First, an original genomic relationship matrix ($\mathbf{G}$) was built according to VanRaden (2008) and Hayes *et al*. (2009),

$$\mathbf{G} = \mathbf{MM}'/\sum 2p_iq_i$$

where elements in column i of $\mathbf{M}$ are $0-2p_i$, $1-2p_i$ and $2-2p_i$ for genotypes A_1A_1, A_1A_2 and A_2A_2, respectively, and q_i is allele frequency of A_1 and p_i is allele frequency of A_2. In this study, allele frequencies were calculated from the current marker data. $\mathbf{G}$ matrix used in each analysis was built using the marker data of the animals in the corresponding reference and validation data. Thus, the marker data used in construction of $\mathbf{G}$ matrix differed among different data sets. Second, the $\mathbf{G}$ matrix was adjusted to be on the same scale as the pedigree-based relationship matrix according to Christensen *et al*. (2012). Thus, the adjusted matrix ($\mathbf{G_c}$) was

$$\mathbf{G_c} = \alpha + \mathbf{G}\beta$$

The parameter α and β were derived from the following equations,

$$\text{Average diagonal}(\mathbf{A}) = \alpha + \text{Average diagonal}(\mathbf{G})\beta$$

$$\text{Average off-diagonal}(\mathbf{A}) = \alpha + \text{Average off-diagonal}(\mathbf{G})\beta$$

where $\mathbf{A}$ was the pedigree-based relationship matrix for the genotyped animals and was extracted from the relationship matrix built based on the whole pedigree. Furthermore, matrix $\mathbf{G_A}$ was calculated as

$$\mathbf{G_A} = (1-\omega)\mathbf{G_c} + \omega\mathbf{A}$$

where ω was the relative weight on matrix $\mathbf{A}$. In this study, $\omega = 0.20$ was chosen according to the previous studies in

Table 1 *Number of animals (dependent on traits) in different validation sets and reference sets*

	Reference set			
Validation set	Reference	Number of DK bulls	Number of US bulls	Number of cows
Validation on bulls	DK bulls	996 to 1028		
211 to 338 bulls	DK + US bulls	996 to 1028	990 to 1150	
Validation on cows	DK bulls	1212 to 1278		
2571 to 3287 cows	DK + US bulls	1212 to 1278	988 to 1148	
	DK bulls + cows	1212 to 1278		4168 to 4836
	DK + US bulls + cows	1212 to 1278	988 to 1148	4168 to 4836

https://doi.org/10.1017/S1751731115001792 Published online by Cambridge University Press

Nordic cattle populations (Su *et al.*, 2011 and 2012c; Gao *et al.*, 2012). In this setting, the GBLUP model is equivalent to a GBLUP including a genomic effect and a residual polygenic effect accounting for 80% and 20% of total additive genetic variance, respectively.

Genomic predictions were performed using the DMU package (Madsen *et al.*, 2010). Additive genetic and residual variances applied in Nordic routine genetic evaluation were used in this study to predict genomic breeding values.

Measurement of consistency in genome and genetic relationship between Danish and US Jersey populations

Gain in genomic prediction accuracy from a joint reference population depends on genetic similarity and relationship between the populations involved. In this study, the consistency of linkage disequilibrium, allele frequency and genetic relationship between Danish and US Jersey populations were investigated. The consistency of linkage disequilibrium was measured as the correlation of r values for adjacent marker pairs between the Danish Jersey bulls and US Jersey bulls, that is, $\mathrm{Cor}(r_{\mathrm{LD(DK)}}, r_{\mathrm{LD(US)}})$, where $r_{\mathrm{LD}} = \frac{f(AB)-f(A)f(B)}{\sqrt{f(A)f(a)f(B)f(b)}}$ for marker A (allele A and a) and marker B (allele B and b). The consistency of marker allele frequency was measured as the correlation of allele frequencies between the two populations. The genetic relationship coefficients between Danish and US Jersey bulls were calculated based on SNP markers. Following Clark *et al.* (2012), the maximum relationship and the mean of top 10 relationships of a Danish bull with the US bulls were used as measures of relationship between a Danish bull and the US bulls.

Results

Reliability of GEBV using Danish bull reference population

The reliabilities of GEBV using reference population comprising Danish bulls alone are shown in Table 2 for validation on bulls and in Table 3 for validation on cows. In validation on bulls, reliability of GEBV for the eight traits ranged from 0.151 (longevity) to 0.372 (milk yield) with an average of 0.262. In validation on cows, reliabilities of GEBV for the six traits ranged from 0.249 (fat) to 0.555 (mastitis) with an average of 0.394. Validation reliabilities on cows were higher than those on bulls for all the traits, except for protein in which reliability for bulls was slightly higher than that for cows. Averaged over the six traits common in the two validations, reliability in validation on cows was 11.7 percentage points higher than that in validation on bulls. There were at least two reasons for higher reliability in validation on cows than bulls. The first could be that the reference population for validation on cows was larger (>200 bulls) than that for validation on bulls. The second was that validation cows were a random sample, while validation bulls are a selected sample (selected on parent average) which would reduce the correlation between GEBV and DRP and thereby result in an underestimate of the true reliability.

As shown in Table 2, the regression coefficients of DRP on GEBV for validation bulls were considerably <1, except for fertility which was slightly >1. The regression coefficients were 1.09 for fertility and ranged from 0.71 to 0.88 for the other five traits. However, the regression coefficients for validation cows (Table 4) were in general slightly larger

Table 2 *Validation reliability (%) of GEBV and regression coefficient of DRP on GEBV using Danish bull reference population (DK) and joint DK-US bull reference population (DKUS), based on validation on bulls*

		r^2_{GEBV}		Regression	
Trait	*n*	DK	DKUS	DK	DKUS
Milk	338	37.2	44.1	0.88	0.83
Fat	338	21.2	22.2	0.71	0.68
Protein	338	29.5	32.9	0.72	0.69
Fertility	271	28.9	27.4	1.09	1.04
Mastitis	299	28.3	28.9	0.73	0.72
Body conformation	275	29.9	34.0	0.83	0.79
Udder conformation	275	20.0	30.2	0.72	0.81
Longevity	211	15.1	14.0	0.71	0.62
Average	293	26.2	29.2	0.80	0.77

GEBV = genomic estimated breeding value; DRP = de-regressed proofs.

Table 3 *Validation reliabilities (%) of GEBV using Danish bull reference population (DK), joint DK-US bull reference population (DKUS), Danish bull and cow reference population (DKCOW), DK-US bull and cow reference population (DKUSCOW), based on validation on cows*

Trait	*n*	DK	DKUS	DKCOW	DKCOWUS
Milk	3287	44.2	53.1	65.8	68.5
Fat	3287	24.9	31.7	36.1	38.2
Protein	3287	28.5	35.9	40.3	42.1
Mastitis	3287	55.5	57.0	56.3	54.9
Body conformation	2572	42.6	49.4	40.7	43.6
Udder conformation	2572	40.6	49.1	46.3	51.8
Average	3049	39.4	46.0	47.6	49.9

GEBV = genomic estimated breeding value.

Table 4 *Regression coefficients of GEBV on DRP for genomic prediction using Danish bull reference population (DK), joint DK-US bull reference population (DKUS), Danish bull and cow reference population (DKCOW), DK-US bull and cow reference population (DKUSCOW), based on validation on cows*

Trait	*n*	DK	DKUS	DKCOW	DKCOWUS
Milk	3287	1.28	1.29	1.23	1.24
Fat	3287	0.80	0.90	0.86	0.88
Protein	3287	0.93	1.00	0.93	0.93
Mastitis	3287	1.19	1.15	1.14	1.06
Body conformation	2572	1.11	1.11	0.99	1.00
Udder conformation	2572	1.25	1.28	1.10	1.13
Average	3049	1.09	1.12	1.04	1.04

GEBV = genomic estimated breeding value; DRP = de-regressed proofs.

https://doi.org/10.1017/S1751731115001792 Published online by Cambridge University Press

than 1, except for protein (0.93) and fat (0.80). The inconsistency in regression coefficients between the two validation sets could reflect difference in correlation coefficients between random sample (cows) and selected sample (bulls). Mäntysaari *et al.* (2010) pointed out that selection of test bulls will reduce the regression coefficient and reliability of GEBV.

Gain in prediction reliability from including US bulls in reference population

Including US Jersey bulls in the reference population resulted in a large increase in accuracy of genomic prediction. In validation on bulls (Table 2), six of the eight traits benefitted from the joint reference population. The gains by including US Jersey bulls in the reference population ranged from 1.0 percentage points for fat to 10.5 percentage points for udder conformation. However, there was a loss in reliability by 1.5 percentage points for fertility and 1.1 percentage points for longevity. A possible reason for the loss of reliability could be that the definitions of the two traits were not the same in Danish and US Jersey populations. Averaged over all eight traits, reliability of GEBV using the joint reference population was 3.0 percentage points higher than the reliability of GEBV using the Danish bull reference population alone.

The gain by including about 1100 US Jersey bulls in the reference population was more pronounced in validation on cows. As shown in Table 3, reliability of GEBV using the joint DK-US bull reference population was higher than using the Danish bull reference population alone for all six traits. The gain was >6.0 percentage points, except for mastitis which gained 1.5 percentage points, leading to an average gain of 6.6 percentage points. For these six traits the average gain in validation reliability on bulls was 4.4 percentage points. On the other hand, the joint reference population did not reduce bias of GEBV. The regression coefficients of DRP on GEBV were in general similar to those when using the Danish bull reference population, that is, generally <1 in validation on bulls (Table 2) and slightly >1 in validation on cows.

Gain in prediction reliability from including cows in the reference population.

Including about 4800 cows in the reference population led to a large increase in reliability of GEBV (Table 3). Compared with genomic predictions using the Danish bull reference population, using the reference population comprising Danish bulls and cows increased the reliabilities of GEBV by >11 percentage points for the three production traits and by 5.7 percentage points for udder conformation. However, there was only a slight increase of reliability for mastitis (0.8 percentage points) and a slight decrease for body conformation (−1.9 percentage points). Averaged over the six traits in validation on cows, the increase of reliabilities was 8.2 percentage points.

When the reference population already included US bulls, the further gain from including cows became relatively smaller. There was still large gain for the three production traits but no gain for mastitis and body conformation. The gain averaged over the six traits was 3.9 percentage points. Similarly, when the reference population already included cows and Danish bulls, the further gain from including the US bulls was reduced also, leading to an average gain of 2.3 percentage points. Consequently the gain from adding both cows and US bulls, averaged over the six traits, was 10.5 percentage points.

Including cows in the reference population had a small influence on the unbiasedness of GEBV. The regression coefficients of DRP on GEBV from the reference population including cows ranged from 0.80 for fat to 1.28 for milk, while the regression coefficients from the Danish bull reference population were between 0.86 for fat and 1.23 for milk. For all six traits, the regression coefficients for the former scenario deviated slightly less from 1, compared to those for the latter scenario. On the average, the absolute deviation from 1 was 0.115 when using the reference population including cows, while 0.183 when using the Danish bull reference population. These results indicated that the reference population including cows led to a slight improvement in unbiasedness of genomic predictions.

Consistency in genome and genetic relationship between Danish and US Jersey populations

As shown in Table 5, there was a high consistency between Danish and US Jersey populations. The two populations had a similar degree of linkage disequilibrium, a high correlation in linkage disequilibrium up to 0.94, and a high correlation of allele frequency up to 0.91. The accumulative frequency of maximum genomic relationship coefficient and the average of top 10 relationship coefficients between a bull in validation data and US bulls are shown in Figures 1 and 2. The maximum relationship between a Danish validation bull and US bulls ranged from 0.1 to 0.54 with a median of 0.22, that

Table 5 *Degree of linkage disequilibrium (LD) between adjacent SNPs for Danish and US Jersey populations, correlation of LD and correlation of allele frequency between populations*

Population	r^2_{LD}	Cor($r_{LD(DK)}$, $r_{LD(US)}$)	Cor($p_{LD(DK)}$, $p_{LD(US)}$)
DK Jersey	0.260	0.941	0.918
US Jersey	0.271		

SNP = single nucleotide polymorphism.

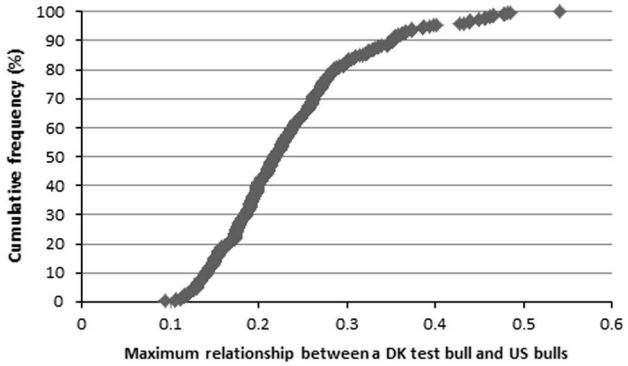

Figure 1 Cumulative frequency against maximum relationship coefficient between a Danish validation bull and the US reference bulls.

https://doi.org/10.1017/S1751731115001792 Published online by Cambridge University Press

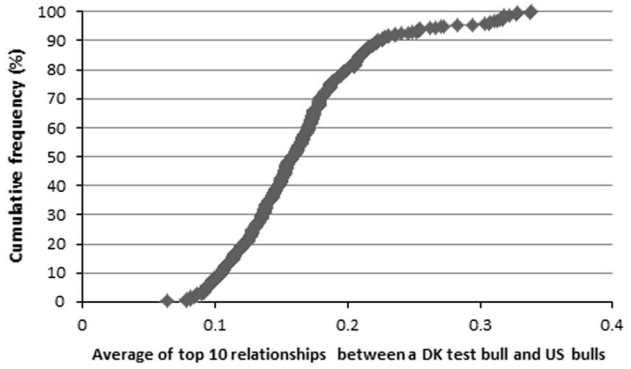

Figure 2 Cumulative frequency against average of top 10 relationship coefficients between a Danish validation bull and the US reference bulls.

is, 50% of Danish validation bulls had a maximum relationship equal to or over 0.22 with one or more US bulls. The median is near the relationship between half-sibs. Correspondingly, the averages of top 10 relationships range from 0.06 to 0.34 with a median 0.16, that is, 50% of Danish test bulls had an average relationship with the closest 10 US bulls equal to or over 0.16. This is equivalent to half of the Danish validation bulls having >10 cousins in the US reference data.

Discussion

This study investigated the gain in reliability of genomic prediction by sharing reference data and adding cows to the reference population. In validation of eight traits on Danish bulls, a joint reference population combining Danish and US reference bulls led to an average increase in reliability of 3 percentage points, compared with genomic prediction using the Danish bull reference population alone. In validation of six traits on Danish cows, the average gain from US reference bulls was 6.6 percentage points, and the average gain from inclusion of cows in reference population was 8.2 percentage points. However, the gains from US bulls and from cows were not accumulative, and the total gain was 10.5 percentage points in the validation on cows. There are at least two reasons that can cause non-accumulative gains in reliability. One is that the information sources (cow information and US bull information here) are not independent (Harris and Johnson, 1998), and the other is that the increase of reliability with increase of reference population size is not linear (Daetwyler *et al.*, 2008; Goddard, 2009; Goddard and Hayes, 2009).

Improving genetic prediction by sharing reference data

Sharing reference data is an efficient approach to increase the size of a reference population and consequently improve the accuracy of genomic predictions. Previous studies have reported that the reliability of genomic prediction can be increased by using a joint reference population combining reference animals from other populations. The reliabilities of GEBV increased by 10 percentage points when using a joint reference data combining reference bulls of four European Holstein populations, compared with those obtained from national reference population alone (Lund *et al.*, 2011). A large improvement was also realized when combining Holstein populations in North America (Schenkel *et al.*, 2009; Muir *et al.*, 2010). Reliability was 3.2 percentage points higher for Brown Swiss cattle using a joint reference population including foreign bulls in the US domestic prediction (VanRaden *et al.*, 2012).

Inclusion of about 1150 US Jersey bulls resulted in a large increase in accuracy of genomic prediction. The results were consistent with the validation of Danish bulls in US scale (i.e., performance in US) using joint US-Danish reference population (Wiggans *et al.*, 2015). At least two reasons can explain this large gain. First, the Danish reference population was small; the inclusion of US Jersey bulls doubled the size of reference population. Large benefit from including reference animals of another population to a small national reference population has been reported by Zhou *et al.* (2013), who added 4400 Danish progeny-tested Holstein bulls to the Chinese reference population, which comprised 1500 Chinese Holstein cows. The gain from the inclusion of Danish bulls was 29 percentage points for Chinese Holstein bulls and 7 percentage points for Chinese Holstein cows.

The second reason for the large gain in prediction accuracy from US Jersey was that there was a high consistency in genome and a strong genetic link between Danish Jersey and US Jersey. Semen of US Jerseys has been used in the Danish Jersey population for a long time, especially during the period from 1985 to 1995. Today the US Jersey breed proportion is about 38% in the Danish Jersey population (http://www.vikinggenetics.com.au/breeds/viking-jersey/about-viking-jersey). Consequently, the correlation of linkage disequilibrium between the two populations was high up to 0.94, and the correlation of allele frequency was 0.91. Furthermore, half of the Danish validation bulls have a relationship coefficient of at least 0.22 with one or more US reference bulls. The importance of relationship between populations for genomic prediction across populations has been reported in many previous studies (Brondum *et al.*, 2011; Zhou *et al.*, 2013 and 2014b).

Improving genetic prediction by including cows in reference population

In dairy cattle the reference population for genomic prediction usually consisted of progeny-tested bulls since they have a large group of daughters with records, thus reliable phenotypic information. However, the phenotypic information of cows is also helpful. Though cow information can be summarized as daughter group means or DRP of sires which are usually in the reference population, this could result in a loss of information from the variation between daughters. Previous studies have reported that a cow reference population leads to moderate reliability of genomic predictions (Ding *et al.*, 2013; Li *et al.*, 2014). A more common approach is to add genotyped cows to the bull

https://doi.org/10.1017/S1751731115001792 Published online by Cambridge University Press

reference population. Calus *et al.* (2013) investigated accuracy of genomic prediction using 1609 cows and 296 bulls as reference animals, and reported that the combined cow and bull reference population resulted in a prediction accuracy higher than using cow reference population alone and much higher than using bull reference population alone. Cooper *et al.* (2015) reported that adding 30 852 cows to the bull reference population (21 833 bulls) increased reliability by 0.4 percentage points for validation bulls and 4.4 points for validation cows. In a simulation study with 60 progeny-tested bulls and 2000 or 4000 cows as reference population, the reliability of genomic prediction using the combined bull and cow reference population was nearly twice as high as using the bull reference population alone (Thomasen *et al.*, 2014). The large increase of prediction accuracy by including cows in reference population was also reported in another simulation study (Buch *et al.*, 2012).

In the current study, inclusion of about 4800 cows in the reference population increased reliability of GEBV by 8.2 percentage points. It can be argued that the gain may be overestimated, because most cows in the reference population were contemporaries of the validation cows, which may be favorable for the prediction of the validation cows. However, in this study all half and full sibs of the validation cows as well as their offspring were excluded from the reference population. Therefore, the overestimation of the gain from cows is not expected to be an issue. The large gain from inclusion of cows in the reference population could be due to the fact that the bull reference population was small (about 1250 bulls). The inclusion of cows actually greatly increased the size of the reference population. When the reference population already included US bulls, the further gain from cows decreased to 3.9 percentage points. This indicates that inclusion of cows in the reference population greatly benefits populations with a small reference data set, but may not necessary largely benefit populations with large reference data sets. It should be pointed that only half of the cows available were used as reference animals in this study. The remaining cows were used as either in the validation set or deleted because they were the close relatives of the validation cows. In practical genomic evaluation a larger gain from including cow information would be obtained since all these cows can be used as reference animals.

Some previous studies have detected bias of genomic prediction when including cows in reference population (Wiggans *et al.*, 2010 and 2011; Dassonneville *et al.*, 2012a). However, in the current study, inclusion of cows in reference population actually slightly reduced bias of GEBV. This is due to the fact that most cows in the analysis were from herds with good data registration where all cows available were genotyped. Therefore, bias due to preferential treatment of bull's dam is not an issue in current study.

Many countries have genotyped cows either to increase the size of reference population or to select females or bull dams. To reduce the cost of genotyping, cows are usually genotyped with a low density chip. Previous studies have reported that the accuracy of imputation from low density panel (7k) to Bovine SNP50 panel (54k) is over 97% (Boichard *et al.*, 2012; Dassonneville *et al.*, 2012b; Su *et al.*, 2014). This indicates that genotyping cows for genomic prediction is feasible.

Alternative approaches to improve genomic prediction for small breeds

In addition to sharing reference data and including cows in the reference population, there are many alternative approaches that may improve accuracy of genomic prediction for numerically small breeds. One approach is to use a single-step model for genomic prediction (Legarra *et al.*, 2009; Aguilar *et al.*, 2010; Christensen and Lund, 2010). Single-step models have the advantage that they directly use information of both genotyped and non-genotyped animals by integrating genomic, pedigree and phenotype information in a single-step procedure. Makgahlela *et al.* (2014) predicted GEBV using a single-step model in which DRP of all cows in the Nordic Red population were used as response variables. It allowed using information of all animals, especially directly using dam information to predict breeding value of an individual. The single-step approach increased reliability by 5 to 8 percentage points for yield traits, compared with a GBLUP model using only DRP of genotyped bulls as the response variable (Makgahlela *et al.*, 2013).

Another alternative is to use a multi-breed reference population that combines information from numerically large breeds. However, previous studies have reported that multi-breed reference population can improve reliability of genomic prediction if the breeds involved have a genetic link (Brondum *et al.*, 2011; Zhou *et al.*, 2014a), and very little effect on accuracy of genomic prediction for the genetically distant breeds (Karoui *et al.*, 2012; Zhou *et al.*, 2014b). One of the reasons that no or very little gain is observed from using multi-breed genomic prediction for genetically distant breeds could be due to differences in linkage disequilibrium between breeds. A possible solution could be to detect causal variants based on sequence data. This would eliminate the reliance on linkage disequilibrium, and thus the information of other breeds can be efficiently used for genomic prediction through the covariance structure of the detected causal variants.

Conclusions

Both sharing reference data and including cows in the reference population greatly increased reliability of genomic prediction in Danish Jersey. The gain in reliability of GEBV from the two approaches was >10 percentage points. The results indicate that sharing reference data and including cows in the reference population are efficient approaches to increase reliabilities of genomic predictions and thus increase genetic gain, especially for populations where the number of progeny-tested bulls is small. Therefore, by efficiently using information recourses, genomic prediction for numerical small breeds is promising.

https://doi.org/10.1017/S1751731115001792 Published online by Cambridge University Press

Acknowledgment

This work was performed within the project "Genomic in herds", funded by VikingGenetics and Nordic Cattle Genetic Evaluation.

References

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S and Lawlor TJ 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93, 743–752.

Boichard D, Chung H, Dassonneville R, David X, Eggen A, Fritz S, Gietzen KJ, Hayes BJ, Lawley CT, Sonstegard TS, Van Tassell CP, VanRaden PM, Viaud-Martinez KA, Wiggans GR and Bovine LDC 2012. Design of a bovine low-density SNP array optimized for imputation. PLoS One 7, e34130.

Brondum RF, Rius-Vilarrasa E, Stranden I, Su G, Guldbrandtsen B, Fikse WF and Lund MS 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. Journal of Dairy Science 94, 4700–4707.

Buch LH, Kargo M, Berg P, Lassen J and Sorensen AC 2012. The value of cows in reference populations for genomic selection of new functional traits. Animal 6, 880–886.

Calus MPL, de Haas Y and Veerkamp RF 2013. Combining cow and bull reference populations to increase accuracy of genomic prediction and genome-wide association studies. Journal of Dairy Science 96, 6703–6715.

Christensen OF and Lund MS 2010. Genomic prediction when some animals are not genotyped. Genetics Selection Evolution 42, 2.

Christensen OF, Madsen P, Nielsen B, Ostersen T and Su G 2012. Single-step methods for genomic evaluation in pigs. Animal 6, 1565–1571.

Clark SA, Hickey JM, Daetwyler HD and van der Werf JHJ 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genetics Selection Evolution 44, 4.

Cooper TA, Wiggans GR and VanRaden PM 2015. Short communication: analysis of genomic predictor population for Holstein dairy cattle in the United States-Effects of sex and age. Journal of Dairy Science 98, 2785–2788.

Daetwyler HD, Villanueva B and Woolliams JA 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3, e3395.

Dassonneville R, Baur A, Fritz S, Boichard D and Ducrocq V 2012a. Inclusion of cow records in genomic evaluations and impact on bias due to preferential treatment. Genetics Selection Evolution 44, 40.

Dassonneville R, Fritz S, Ducrocq V and Boichard D 2012b. Short communication: imputation performances of 3 low-density marker panels in beef and dairy cattle. Journal of Dairy Science 95, 4136–4140.

Ding X, Zhang Z, Li X, Wang S, Wu X, Sun D, Yu Y, Liu J, Wang Y, Zhang Y, Zhang S, Zhang Y and Zhang Q 2013. Accuracy of genomic prediction for milk production traits in the Chinese Holstein population using a reference population consisting of cows. Journal of Dairy Science 96, 5315–5323.

Gao HD, Christensen OF, Madsen P, Nielsen US, Zhang Y, Lund MS and Su G 2012. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. Genetics Selection Evolution 44, 8.

Goddard M 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136, 245–257.

Goddard ME and Hayes BJ 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics 10, 381–391.

Habier D, Tetens J, Seefried FR, Lichtner P and Thaller G 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution 42, 5.

Harris B and Johnson D 1998. Approximate reliability of genetic evaluations under an animal model. Journal of Dairy Science 81, 2723–2728.

Hayes BJ, Visscher PM and Goddard ME 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics Research 91, 47–60.

Karoui S, Jesus Carabano M, Diaz C and Legarra A 2012. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genetics Selection Evolution 44, 39.

Legarra A, Aguilar I and Misztal I 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92, 4656–4663.

Li X, Wang S, Huang J, Li L, Zhang Q and Ding X 2014. Improving the accuracy of genomic prediction in Chinese Holstein cattle by using one-step blending. Genetics Selection Evolution 46, 66.

Lund MS, de Ross SP, de Vries AG, Druet T, Ducrocq V, Fritz S, Guillaume F, Guldbrandtsen B, Liu Z, Reents R, Schrooten C, Seefried F and Su G 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genetics Selection Evolution 43, 43.

Lund MS, Su G, Janss L, Guldbrandtsen B and Brondurn RF 2014. Invited review: genomic evaluation of cattle in a multi-breed context. Livestock Science 166, 101–110.

Lund MS, Su G, Nielsen US and Aamand GP 2009. Relation between accuracies of genomic predictions and ancestral links to the training data. In Proc. Interbull Bulletin, Barcelona, Spain, 162–166pp..

Madsen P, Su G, Labouriau R and Christensen OF 2010. DMU –- A package for analyzing multivariate mixed models. I CD communication – Proceeding of the 9th WCGALP, August 1–6, paper No. 732, Leipzig, Germany.

Makgahlela ML, Stranden I, Nielsen US, Sillanpaa MJ and Mantysaari EA 2013. The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations. Journal of Dairy Science 96, 5364–5375.

Makgahlela ML, Stranden I, Nielsen US, Sillanpaa MJ and Mantysaari EA 2014. Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a multibreed population. Journal of Dairy Science 97, 1117–1127.

Mäntysaari E, Liu Z and VanRaden PM 2010. Interbull validation test for genomic evaluations. Interbull Bulletin 41, 17–22.

Muir B, Van Doormaal B and Kistemaker G 2010. International Genomic Cooperation – North American perspective. Interbull Bulletin 41, 71–76.

Pszczola M, Strabel T, Mulder HA and Calus MPL 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. Journal of Dairy Science 95, 389–400.

Sargolzaei M, Chesnais JP and Schenkel FS 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15, 478.

Schenkel FS, Sargolzaei M, Kistemaker G, Jansen GB, Sullivan P, Van Doormaal BJ, VanRaden PM and Wiggans GR 2009. Reliability of genomic evaluation of Holstein cattle in Canada. Interbull Bulletin 39, 51–57.

Su G, Brondum RF, Ma P, Guldbrandtsen B, Aamand GR and Lund MS 2012a. Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. Journal of Dairy Science 95, 4657–4665.

Su G, Christensen OF, Ostersen T, Henryon M and Lund MS 2012b. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS One 7, e45293.

Su G, Madsen P, Nielsen US, Mäntysaari EA, Aamand GP, Christensen OF and Lund MS 2012c. Genomic prediction for Nordic Red Cattle using one-step and selection index blending. Journal of Dairy Science 95, 909–917.

Su G, Gao H and Lund MS 2011. Contributions of different sources of information to reliability of genomic prediction for Danish Jersey population. In: Book of Abstracts of the 61th EAAP, 2 August–2 September, Stavanger, pp. 32.

Su G, Guldbrandtsen B, Aamand GP, Stranden I and Lund MS 2014. Genomic relationships based on X chromosome markers and accuracy of genomic predictions with and without X chromosome markers. Genetics Selection Evolution 46, 47.

Thomasen JR, Guldbrandtsen B, Su G, Brondum RF and Lund MS 2012. Reliabilities of genomic estimated breeding values in Danish Jersey. Animal 6, 789–796.

Thomasen JR, Sorensen AC, Lund MS and Guldbrandtsen B 2014. Adding cows to the reference population makes a small dairy population competitive. Journal of Dairy Science 97, 5822–5832.

VanRaden PM 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91, 4414–4423.

VanRaden PM, Olson KM, Null DJ, Sargolzaei M, Winters M and van Kaam JB 2012. Reliability increases from combining 50,000- and 777,000-marker genotypes from four countries. Interbull Bulletin 46, 75–79.

Wiggans GR, Cooper TA and VanRaden PM 2010. Cow adjustments for genomic predictions of Holstein and Jersey bulls. Journal of Dairy Science 93, 533–534.

https://doi.org/10.1017/S1751731115001792 Published online by Cambridge University Press

Wiggans GR, Cooper TA, VanRaden PM and Cole JB 2011. Technical note: adjustment of traditional cow evaluations to improve accuracy of genomic predictions. Journal of Dairy Science 94, 6188–6193.

Wiggans GR, Su G, Cooper TA, Nielsen US, Aamand GP, Guldbrandtsen B, Lund MS and VanRaden PM 2015. Short communication: improving accuracy of Jersey genomic evaluations in the United States and Denmark by sharing reference population bulls. Journal of Dairy Science 98, 3508–3513.

Zhou L, Ding X, Zhang Q, Wang Y, Lund MS and Su G 2013. Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. Genetics Selection Evolution 45, 7.

Zhou L, Heringstad B, Su G, Guldbrandtsen B, Meuwissen T, Svendsen M, Grove H, Nielsen US and Lund MS 2014a. Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. Journal of Dairy Science 97, 4485–4496.

Zhou L, Lund MS, Wang Y and Su G 2014b. Genomic predictions across Nordic Holstein and Nordic Red using the GBLUP model with different genomic relationship matrices. Jounral of Animal Breeding and Genetics 131, 249–257.

https://doi.org/10.1017/S1751731115001792 Published online by Cambridge University Press